Experts in uncertainty
An introduction to the classical model

Tina Nane
Delft University of Technology

TU Delft

The classical model

- The classical method
- Cooke’s model/method
- Cooke’s classical model/method
- Classical model for expert judgement
- Structured expert judgment (SEJ)
- Structured expert judgment: the classical model (wikipedia)
- Structured elicitation of expert judgement(s)

TU Delft

Why EJ?

- No (or not much) data available
- Data sources in an inadequate form for the analysis
- Data sources fraught with problems

(i.e., poor entry, bad data definitions, etc.)

TU Delft

SEJ elicitation

- The process (protocol) of obtaining information about uncertain events from experts

- Variable of interest

When will man land on Mars?

5%______, 50%______, 95%______

SEJ elicitation

- The process (protocol) of obtaining information about an uncertain events (probability distributions) from experts
- Variable of interest

When will man land on Mars?

e_1: 5%__2035__, 50%__2050__, 95%__2070__

e_2: 5%__2018__, 50%__2025__, 95%__2100__

- 50% quantile (50th percentile) – best guess
- 5% quantile (5th percentile) – $P(X \leq 2035) = 0.05$
- 95% quantile (95th percentile) – $P(X \geq 2070) = 0.05$

TU Delft

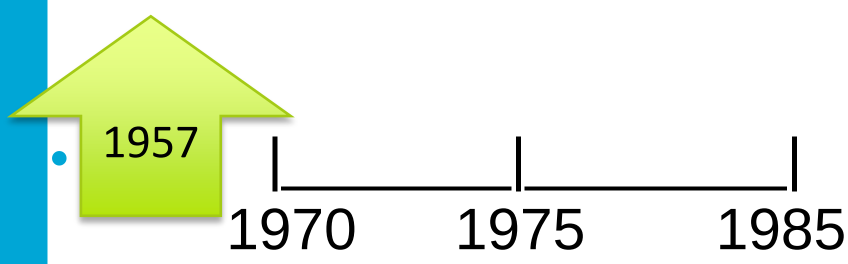
SEJ elicitation

- Questions for which we know the true values (calibration, performance, seed questions/variables)

What was the 1946 RAND forecast for year of first launched satellite?

e_1: 5% __1970__, 50% __1975__, 95% __1985__

e_2: 5% __1950__, 50% __1960__, 95% __2000__

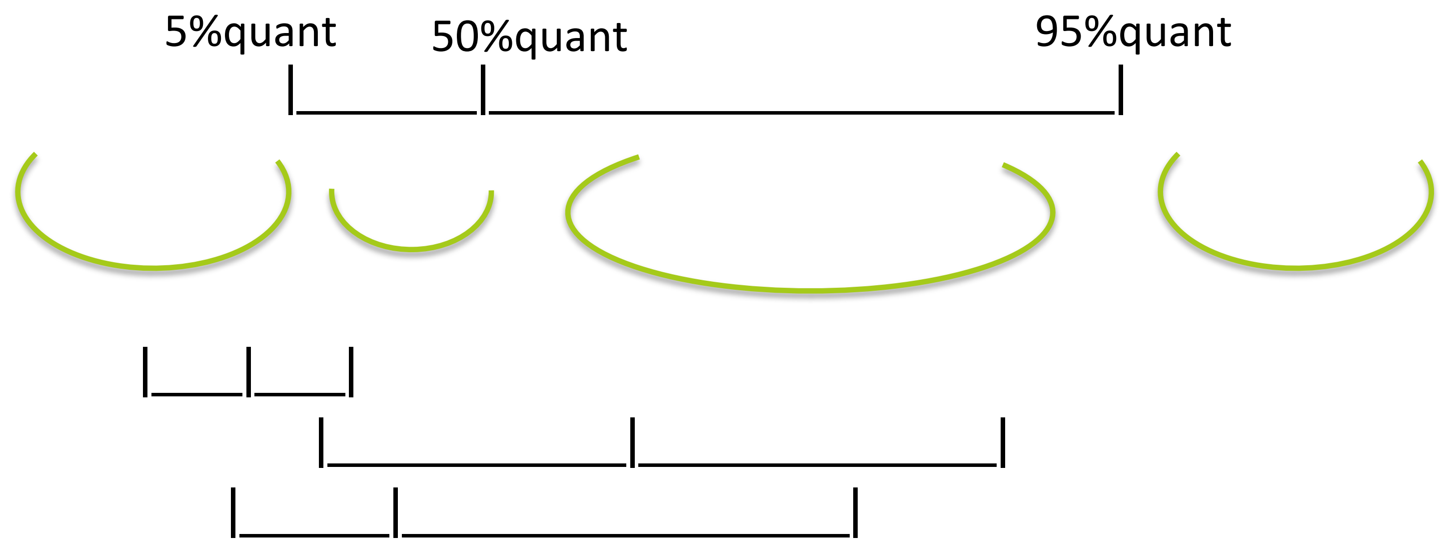


- Sputnik – 4 October, 1957

Calibration

- Are the expert's probability statements statistically accurate?
- Calibration measures how well experts' assessments correspond to actual values (realizations)
- Statistical accuracy
- Calibration is a statistical question – we need a sufficient number of calibration variables (seed questions) to compare assessments to realizations with any confidence

Calibration

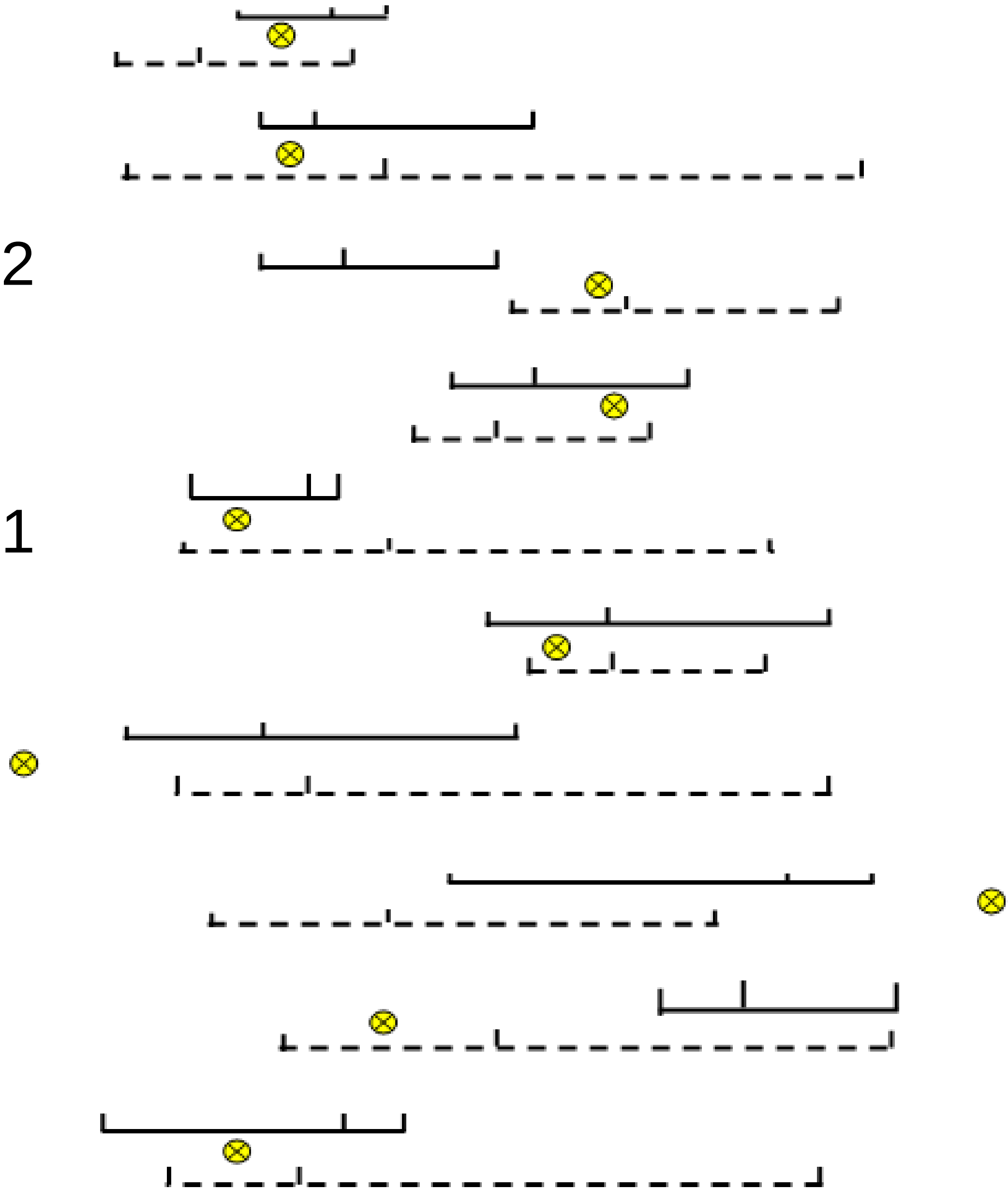


- $p_1 = 0.05$, $p_2 = 0.45$, $p_3 = 0.45$, $p_4 = 0.05$
- $p = (p_1, p_2, p_3, p_4)$
- $s = (s_1, s_2, s_3, s_4)$ for expert e_1?

TU Delft

Calibration

- s_1=0.2, s_2=0.5
  s_3 = 0.1, s_4=0.2
  (for expert 1)
- s_1=0.1, s_2=0.6
  s_3 = 0.2, s_4=0.1
  (for expert 2)



TUDelft

Calibration

- We expect some discrepancy between s and p, due to random fluctuation
- To measure the discrepancy between s and p, we use the relative information of s relative to p

$$I(s,p) = \sum_{i=1}^{4} s_i \ln(s_i/p_i)$$

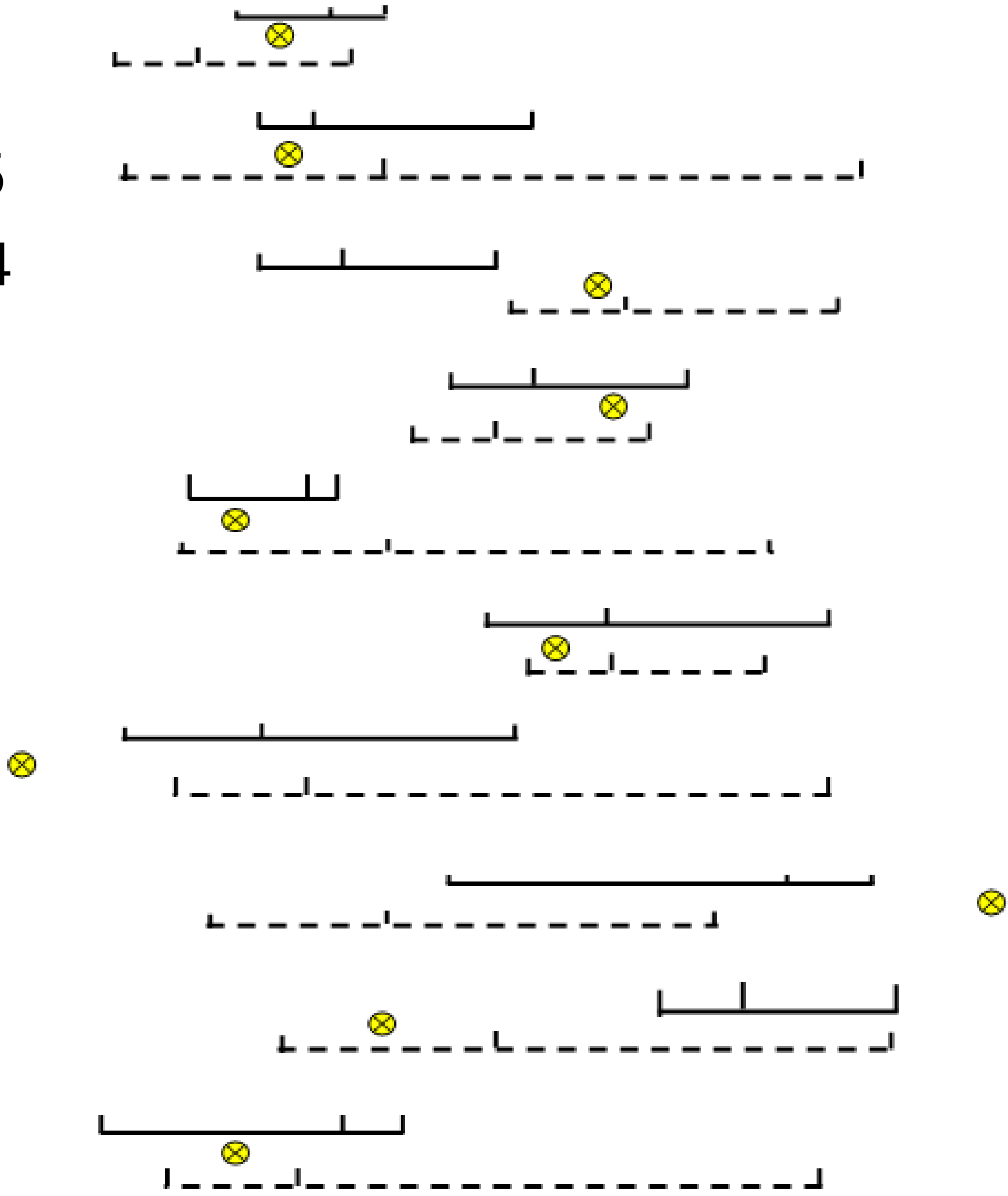
- For n questions and 3 elicited quantiles

$$Cal(e) = 1 - \chi^2_3 (2nI(s,p))$$

TU Delft

Calibration

- Cal(1)=0.0275
  Cal(2)=0.3944



Calibration

Classical model

- Measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate
- We do not say we (fail to) reject expert hypothesis
- Low scores, near zero, mean that it is unlikely that the expert's probabilities are correct
- High scores indicate good support

Information

- Information in a distribution is the degree to which the distribution is concentrated

What was the 1946 RAND forecast for
year of first satellite (1957)?

e_1: 5%__1960__, 50%__1965__, 95%__1970__

e_2: 5%__1950__, 50%__1960__, 95%__2000__

- Measuring information requires associating a density with each assessment of each expert

TUDelft

Information

- Information is measured with respect to a background measure (chosen by the analyst)
  - Uniform
  - Log-uniform
- The background measures require an *intrinsic range* on which measures are concentrated
- k% overshoot rule
  - For each variable, the smallest interval I=[L,U]
  - Extend I to $I^* = [L^*, U^*]$, where

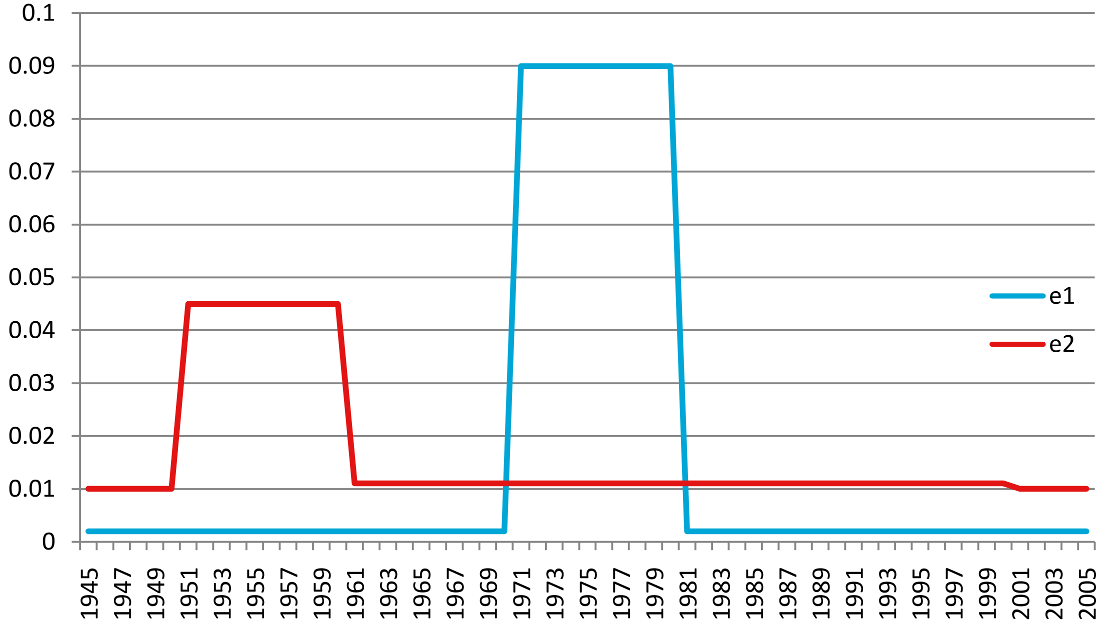
$$L^* = L - \frac{k(U-L)}{100}, U^* = U + \frac{k(U-L)}{100}$$

- k is chosen by the analyst (typically k=10)

TU Delft

Information

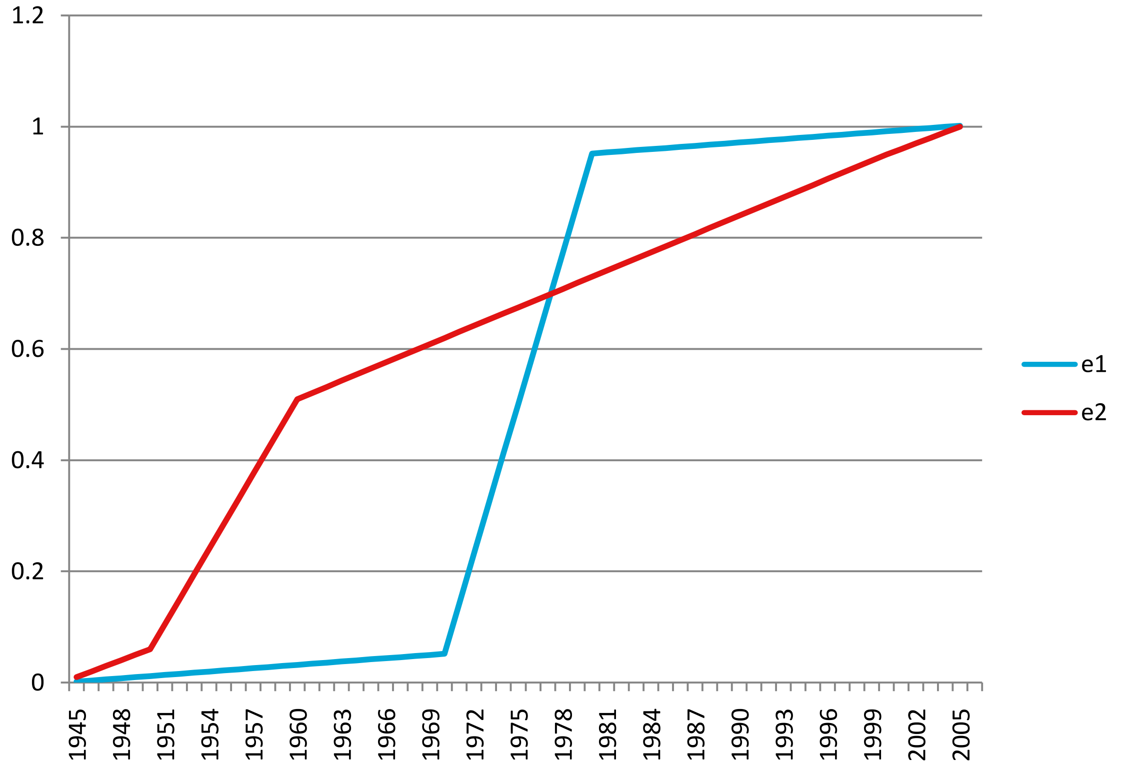
- Probability mass function which agrees with experts' percentile assessments, relative to the uniform measure



- $I^* = [1945, 2005]$

TU Delft

Information



Information

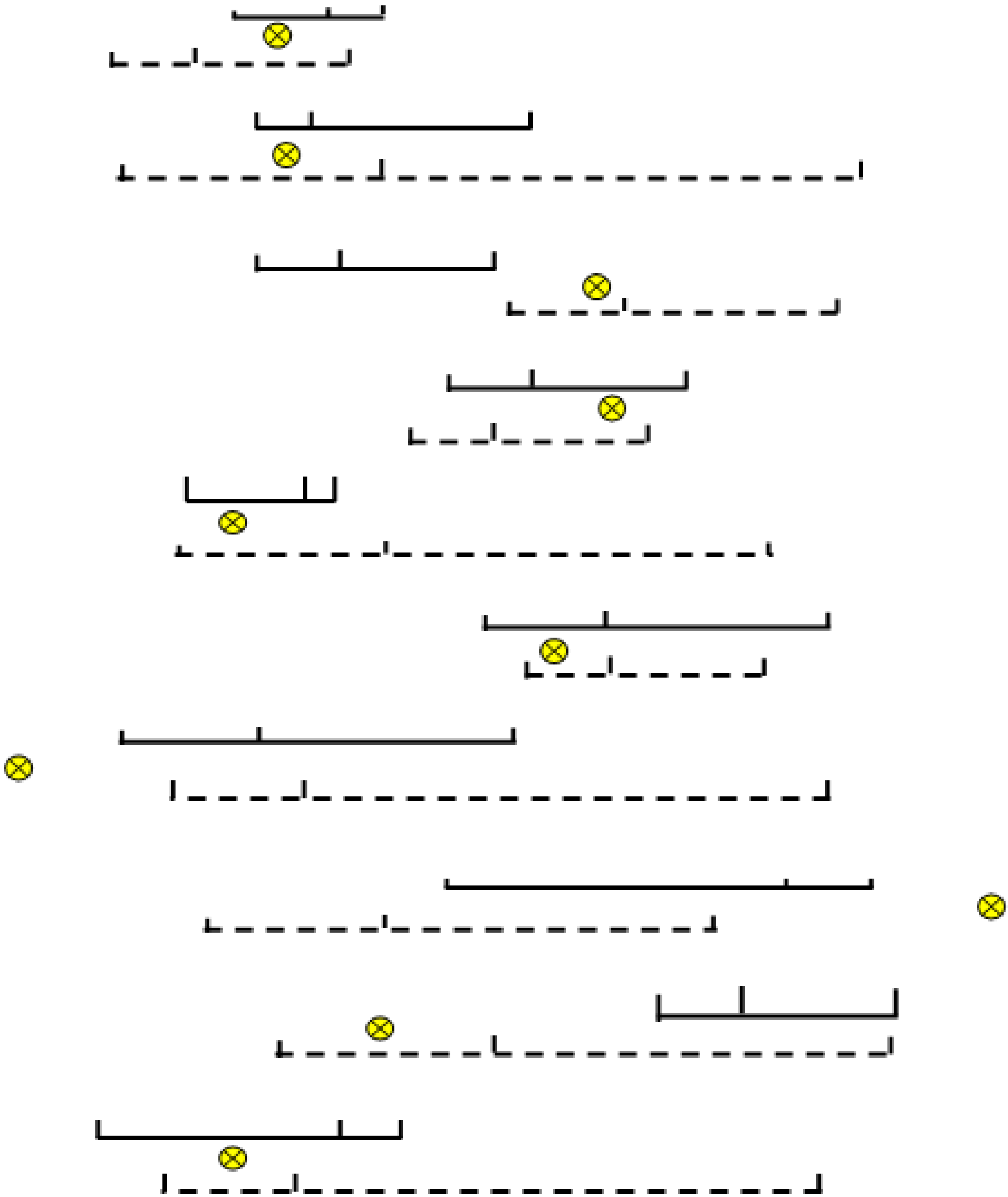
- Inf(i,Unif)= relative information wrt background

$$\mathrm{Inf}(e) = \frac{1}{n}\sum_{i=1}^{n} Inf(f(i), Unif)$$

- High values indicate that the expert is adding "a large amount of information" to the background distribution

Information

- Inf(1)=1.149
  Inf(2)=0.5912



TUDelft

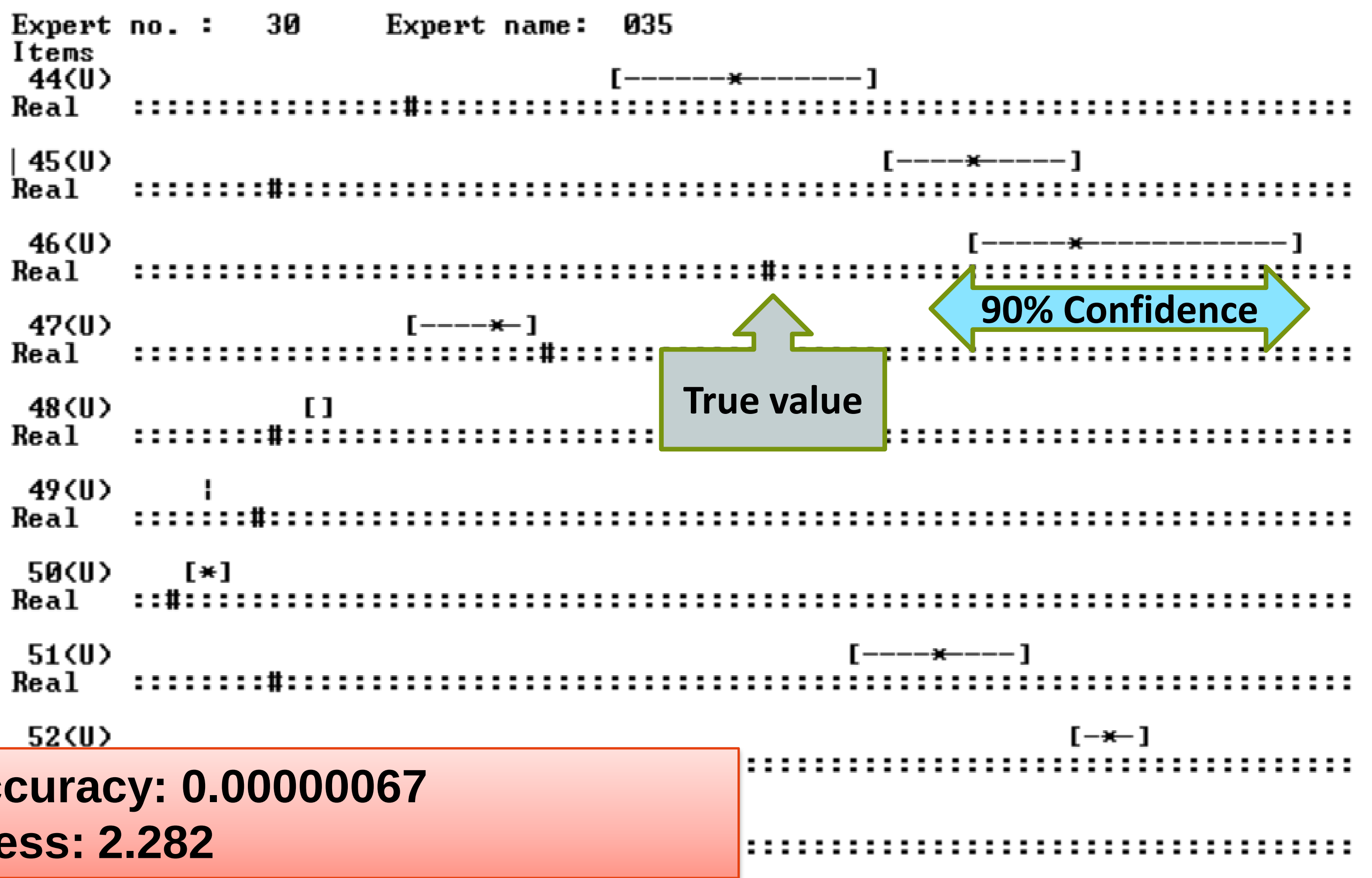
What is a GOOD subjective probability assessor?

- **Calibration (statistical accuracy)**
  - Are expert's probability statements statistically accurate?
  - A high calibration score
- **Informativeness**
  - Probability mass concentrated in a small region, relative to the background measure
  - High information score

TU Delft

Expert performance CAN be objectively measured

Very High Information, Very Poor Statistical Accuracy



Statistical Accuracy: 0.00000067
Informativeness: 2.282

Expert performance CAN be objectively measured

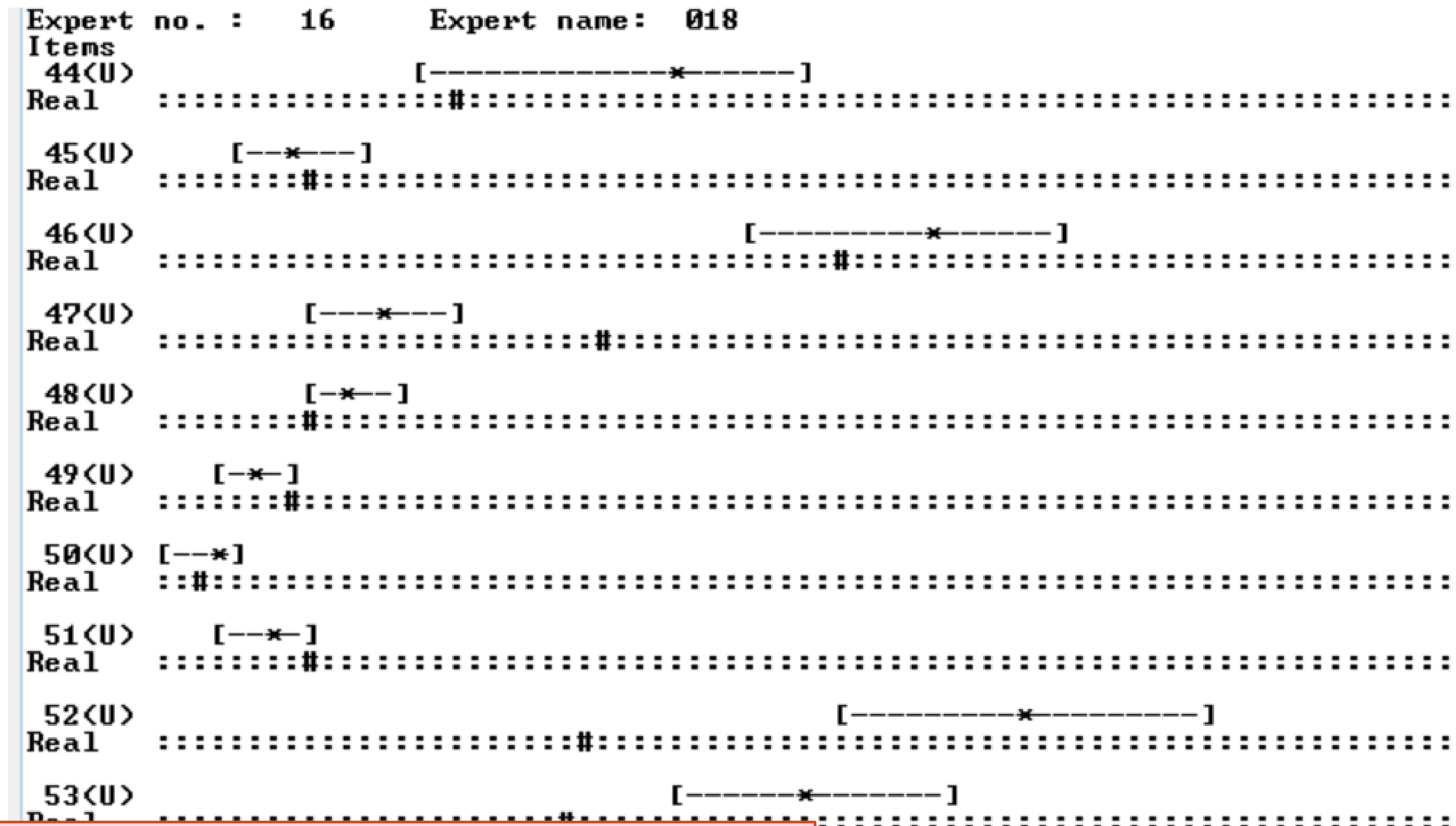
Low Information, Good Statistical Accuracy

```
Expert no. :   29     Expert name:   034
Items
 44(U) [------*----------------------------]
Real   ::::::::::::::::#::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 45(U)   [-----------------------*---------------------------------------]
Real   ::::::::#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 46(U)    [--------------------------*--------------------------------]
Real   ::::::::::::::::::::::::::::::::::::::#::::::::::::::::::::::::::::::::::
 47(U)       [-----*------------------------------]
Real   ::::::::::::::::::::::::#::::::::::::::::::::::::::::::::::::::::::::::::
 48(U)  [----------------------------------------------*-----------------------]
Real   ::::::::#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 49(U) [--------*------------------------------------------]
Real   :::::::#:::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 50(U) [--*-----]
Real   ::#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 51(U)  [----------------*---------------------------------]
Real   ::::::::#::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
 52(U)        [-------------------------------*--------------------------]
Real   ::::::::::::::::::::::::#::::::::::::::::::::::::::::::::::::::::::::::::
```

Statistical Accuracy: 0.57
Informativeness: 0.53

Expert performance CAN be objectively measured

High Information, Decent Statistical Accuracy



Statistical Accuracy: 0.12
Informativeness: 1.77

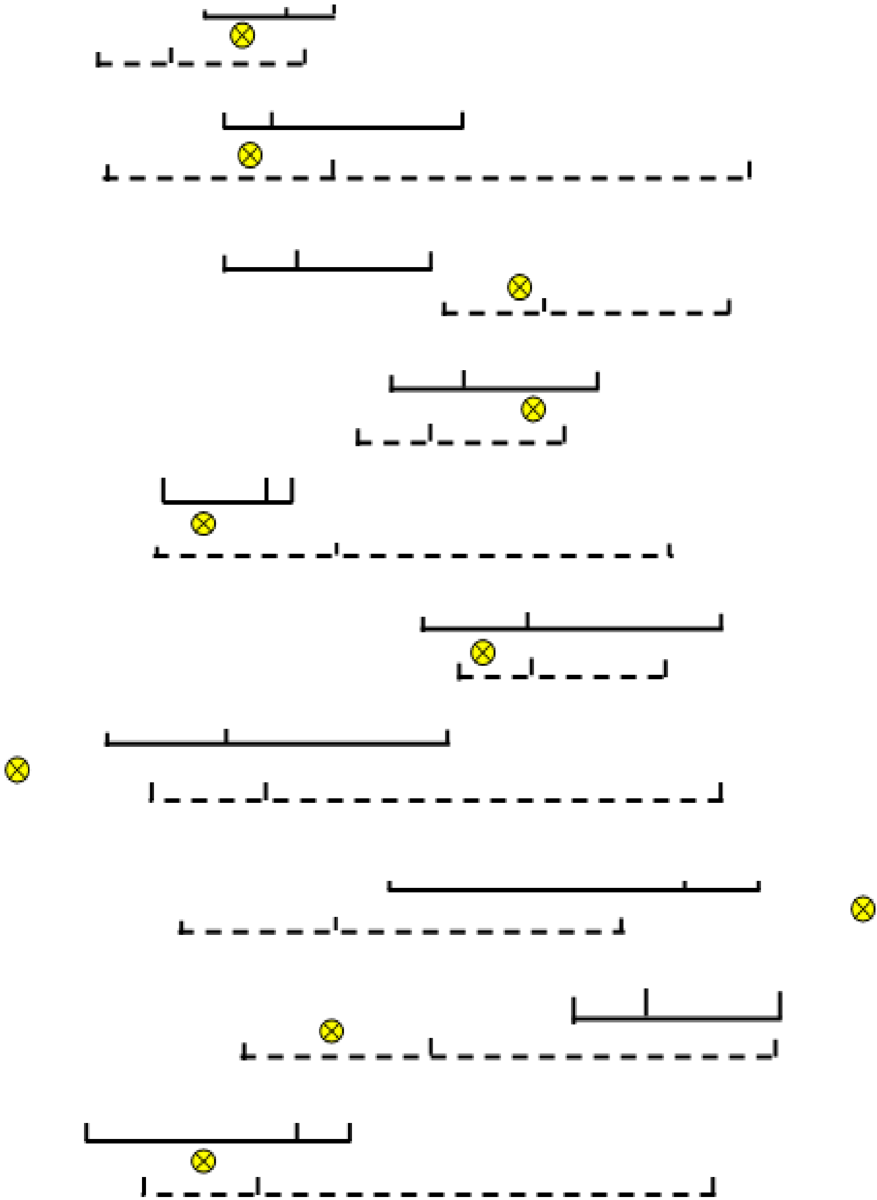
Combined score

- **Cal(e)*Inf(e)**
- Score(1)=

0.0275*1.149=0.0315

- Score(2)=

0.3944*0.5912=0.2331



Decision Maker (DM)

Expert	Calibration	Informativeness
Expert 1	0.06083	0.91
Expert 2	0.00628	1.56
Expert 3	0.01397	1.24
Expert 4	0.6827	0.82
Expert 5	0.002809	1.15
Expert 6	0.05706	1.32
Expert 7	0.01397	1.10

- $Cal(e) \geq 0.05$

TUDelft

Combined expert score

- $w_\alpha(e) = Cal(e) * Inf(e) * \underbrace{1\{Cal(e) \geq \alpha\}}$

Significance Level

= 1 if calibration ≥ α, else = 0

- This score is an asymptotically proper scoring rule, i.e.
  - Expert maximizes long run expected score by, and only by, stating percentiles which (s)he believes

$$DM_\alpha = \frac{\sum_{i=1}^{E} w_\alpha(e_i) f_{e,i}}{\sum_{i=1}^{E} w_\alpha(e_i)}$$

TUDelft

Optimization

-
- Q: How to choose α?
- A: For each α, compute

$$Cal(DM_\alpha) * Inf(DM_\alpha)$$

- choose α for which this is maximum

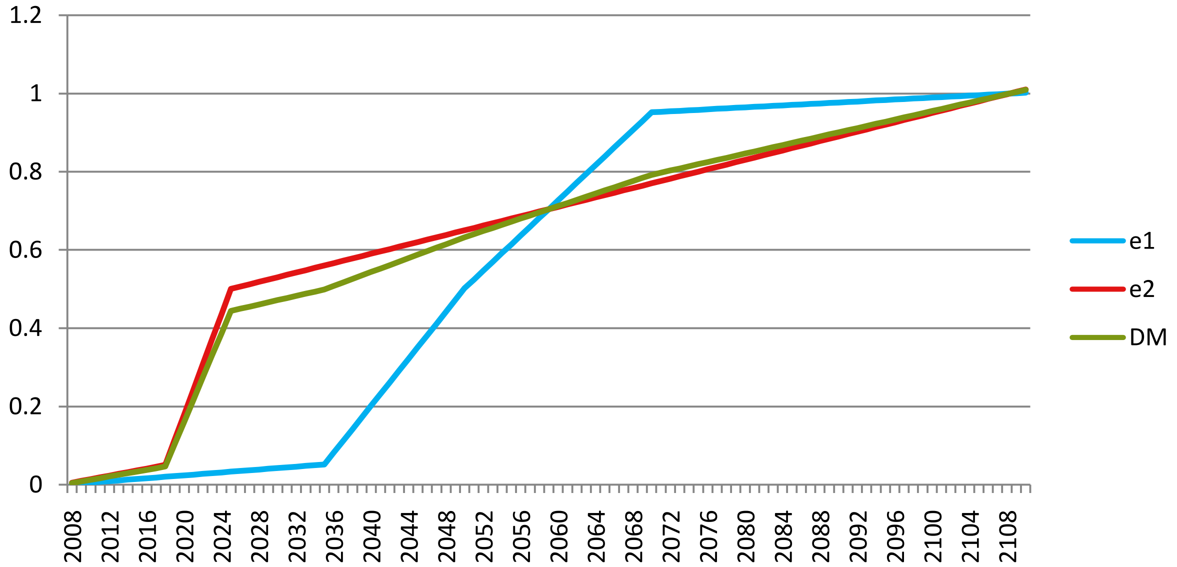
TU Delft

Decision Maker (DM)

- **Performance based weights (PW)**

$$w_1 = \frac{0.0315}{0.0315 + 0.2331} = 0.12$$

$$w_2 = \frac{0.2331}{0.0315 + 0.2331} = 0.88$$



Another approach – Averaging quantiles

- **Performance based weights (PW)**

$$w_1 = 0.12$$

$$w_2 = 0.88$$

- $\boldsymbol{DM_{PW} = w_1 e_1 + w_2 e_2}$
  - 50% - 0.12*2050+0.88*2025=2028
  - 5% - 0.12*2035+0.88*2018=2020
  - 95% - …

TU Delft

Another option - Equal weights

- When will man land on Mars?

  e_1: 5%__2035__, 50%__2050__, 95%__2070__

  e_2: 5%__2018__, 50%__2025__, 95%__2100__

- Equal weights?
  - 50% - $\frac{1}{2}\,2050 + \frac{1}{2}2025 = 2037$
  - 5% ..., 95% ..., f(i)

$$DM_{EW} = \frac{1}{2}e_1 + \frac{1}{2}e_2$$

TUDelft

Combining experts

- Equal weights

Performance based

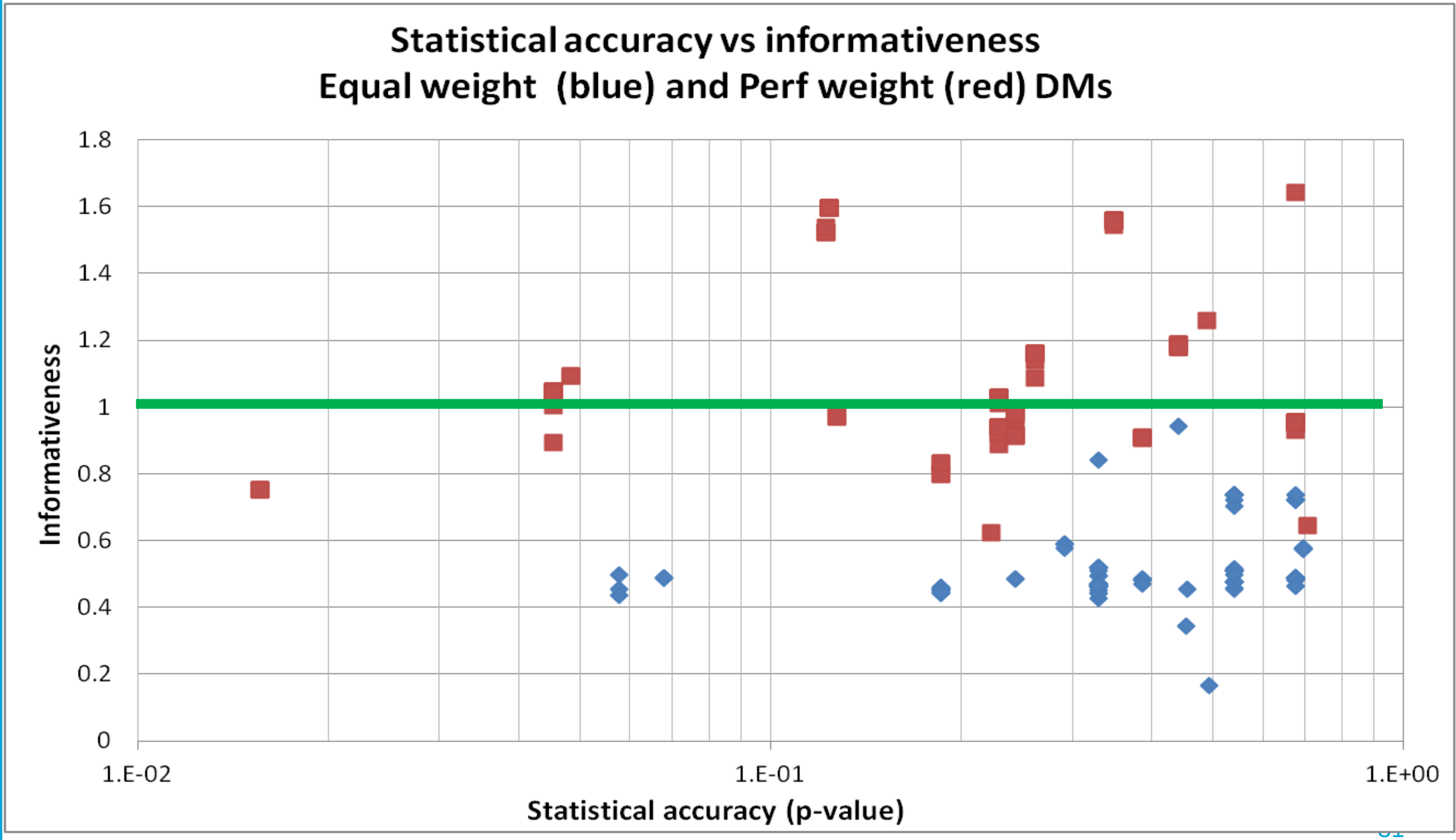
- Global weights: Calibr*Ave Inf*cutoff

  Item weights: Calibr*Inf per item*cutoff

TUDelft

Why bother?

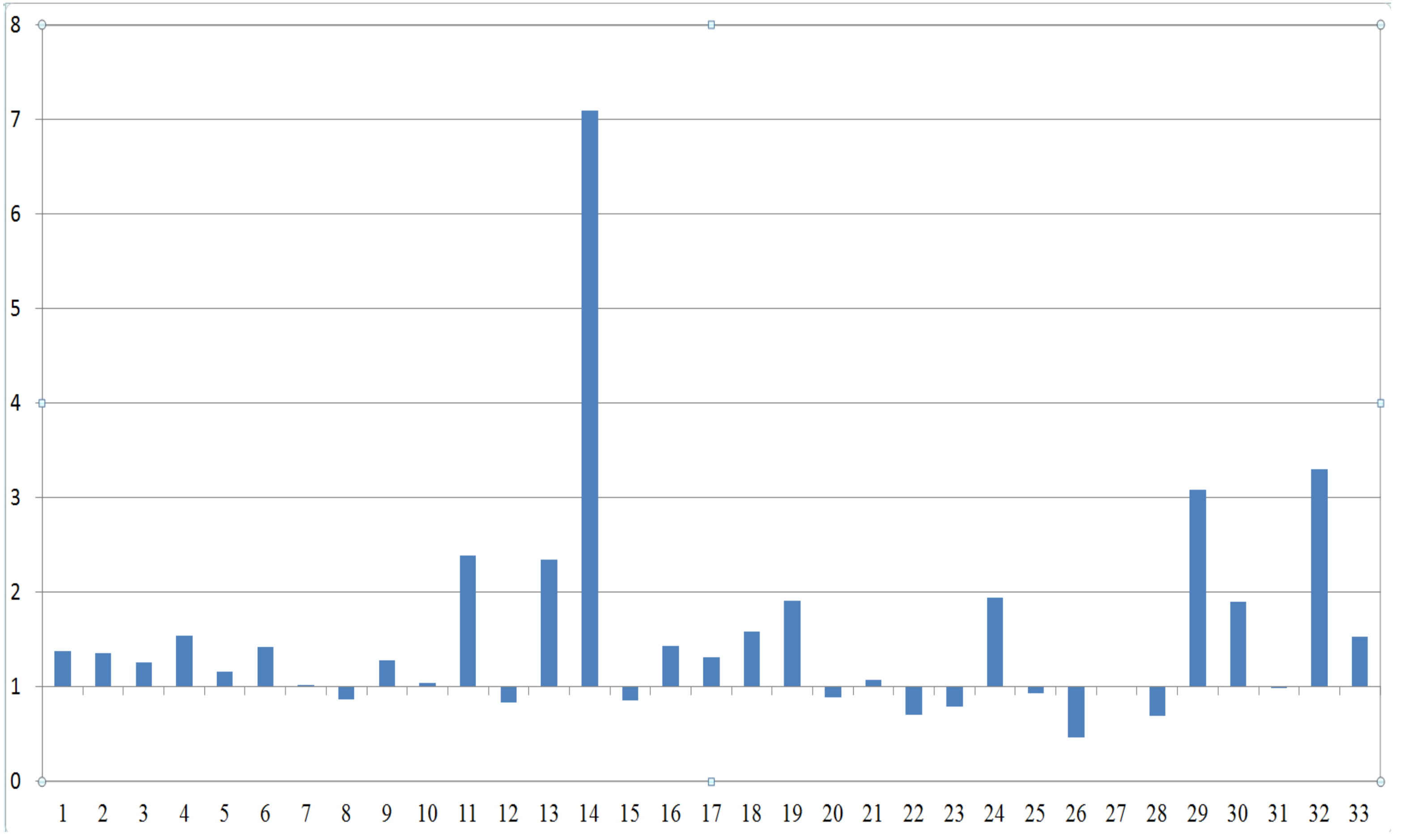
- **Performance based combination of experts improves statistical accuracy and informativeness**



TU Delft

Out of sample validation

- **Training sets sized at 80% of the calibration variables**
- **Ratio PWCombined/EWCombined**



TU Delft

Questions?

TU Delft

SEJ for rational consensus

- Parties pre-commit to a method which satisfies necessary conditions for scientific method
  - Traceability/accountability
  - Neutrality (don't encourage untruthfulness)
  - Fairness (ab initio, all experts are equal)
  - Empirical control (performance measurement)
- Goal: comply with the principles and combine experts' judgements to get a GOOD probability assessor

TU Delft

SEJ theses

- SEJ is not knowledge
- Experts can quantify uncertainty as subjective probability
- Experts don't agree
- We can do better than equal weighting
- *The choice is not whether to use expert judgment, but whether to do it well or badly*

TU Delft